Supplement
Improving Table Structure Recognition with Visual-Alignment Sequential Coordinate Modeling

Yongshuai Huang[*,1] Ning Lu[*,1] Dapeng Chen[1]
Yibo Li[2] Zecheng Xie[1] Shenggao Zhu[1] Liangcai Gao[2] Wei Peng[1]
[1] Huawei Technologies Ltd. [2] Peking University
{huangyongshuai1,luning12,chendapeng8,xiezecheng1,zhushenggao,peng.wei1}@huawei.com
{yiboli,gaoliangcai}@pku.edu.cn

1. Datasets

Among all public accessible datasets, TABLE2LATEX-450K [1], TableBank [3], PubTabNet [8] and FinTabNet [7] are used for logical structure recognition, where their logical structures are represented by markup languages such as HTML or Latex. The others are for physical structure recognition, and their structure is described by the bounding box and logical location of the cells, *i.e.* star-row, end-row, start-column, and end-column. Since VAST predicts the logical structure of the table and bounding box of the content, datasets without content bounding box annotations such as TABLE2LATEX-450K, TableBank, UNLV [6], IC19B2H [2], WTW [4] and TUCD [5] are not suitable for our method.

2. Qualitative Results

We present some positive and negative samples of the detection results of non-empty cells by VAST in Fig. 1. From these qualitative results, we can see that the bounding box predicted by VAST can tightly enclose the contents of the cell. Negative samples consist mainly of tables with over-segmented or over-merged content. There are two reasons for these errors, one is due to the ambiguity of the annotations (samples of FinTabNet and ICDAR2013), and the other is due to the lack of semantic information of cell content (samples of PubTabNet and PubTables-1M). It is worth noting that, even though VAST incorrectly predicts some cells, it takes local visual information into account when predicting cell bounding boxes. Compared with the results of VAST w/o VA in Fig. 5 of the paper, VAST can significantly reduce over-segmented cells.

3. Details of content extraction

The output of VAST are the logical structure (HTML) of the table and the bounding box of all non-empty cells, which is incomplete for evaluation metrics considering the content, so we need to obtain the content of each non-empty cell through simple post-processing. Fig. 2 illstrates the complete pipeline:

1. If there is a PDF file of the table, such as FinTabNet, SciTSR, ICDAR2013 and PubTables, we use PDFMiner[1] to extract the content bounding box and content of each text line within the table area from the PDF document. If there are only images of the table, such as PubTabNet, we use PSENET to detect text lines and MASTER to recognize texts in text lines. For the fairness of the comparison, we use the pretrained PSENET and MASTER model of TableMaster[2].

2. We match text lines with non-empty cells by using the highest IoU and IoU ≥ 0.1.

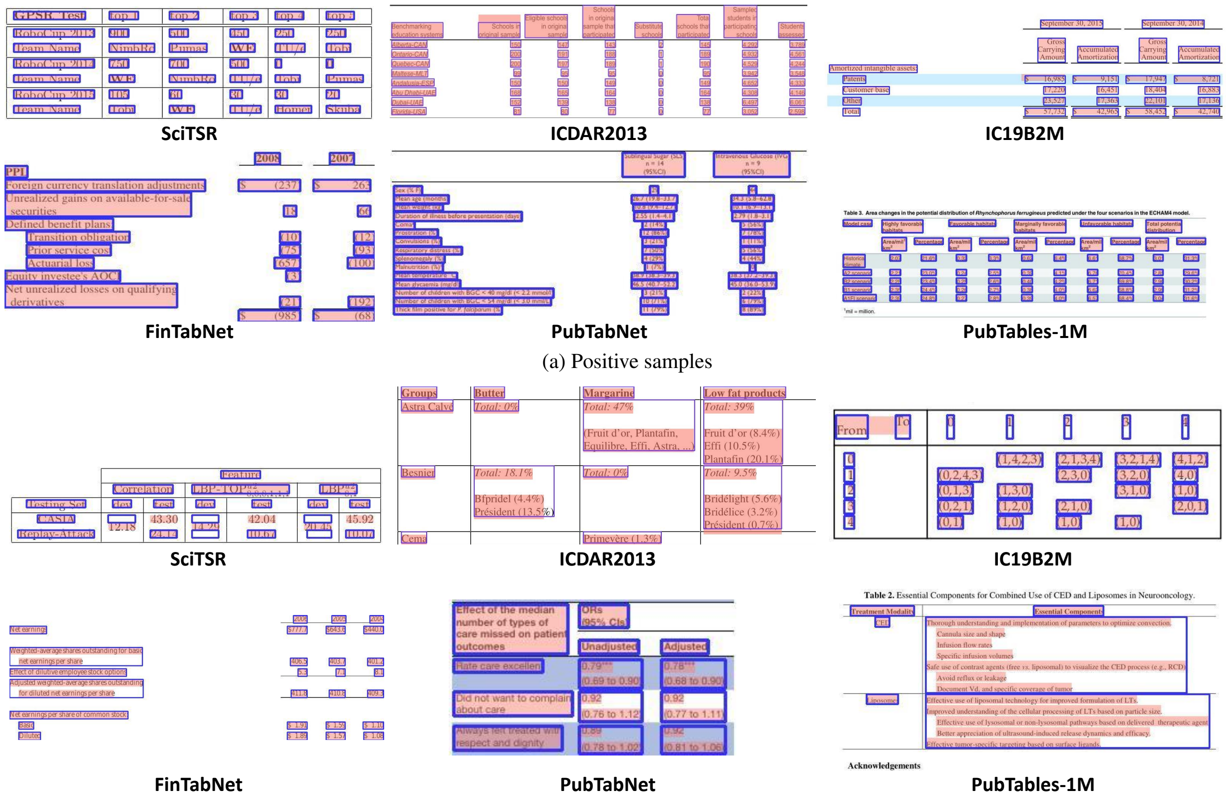
3. For cells that contain multiple text lines, we sort the text lines left-to-right and top-to-bottom then merge their texts.

After getting the contents of non-empty cells, we combine them with structure HTML to output the HTML or XML result for evaluation.

[*] Equal contribution.

[1] https://www.unixuser.org/ euske/python/pdfminer/index.html

[2] https://github.com/JiaquanYe/TableMASTER-mmocr


SciTSR
ICDAR2013
IC19B2M
FinTabNet
PubTabNet
PubTables-1M
SciTSR
ICDAR2013
IC19B2M
FinTabNet
PubTabNet
PubTables-1M

(a) Positive samples

(b) Negative samples

Figure 1. Visualization of non-empty cell detection results of VAST

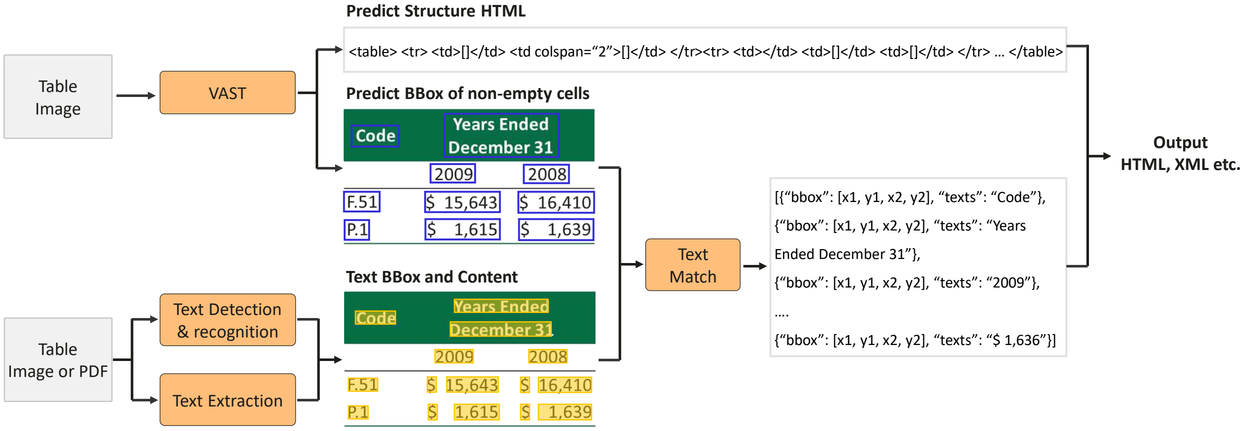

Predict Structure HTML
<table> <tr> <td>[]</td> <td colspan="2">[]</td> </tr><tr> <td></td> <td>[]</td> <td>[]</td> </tr> … </table>
Table Image
VAST
Predict BBox of non-empty cells
Code
Years Ended December 31
2009
2008
F.51
$ 15,643
$ 16,410
P.1
$ 1,615
$ 1,639
Text BBox and Content
Table Image or PDF
Text Detection & recognition
Text Extraction
Text Match
[{"bbox": [x1, y1, x2, y2], "texts": "Code"},
{"bbox": [x1, y1, x2, y2], "texts": "Years Ended December 31"},
{"bbox": [x1, y1, x2, y2], "texts": "2009"},
....
{"bbox": [x1, y1, x2, y2], "texts": "$ 1,636"}]
Output HTML, XML etc.

Figure 2. Pipeline of post-processing to obtain content of non-empty cells

4. Details of cross-attention weight visualization

4.1. Details of the generation of cross-attention visualization maps

In the HTML sequence decoder, we compute the dot-products of the query with all keys, divide each by $\sqrt{d_k}$ and apply a Softmax to get the weight of cross-attention. At the l-th step of decoding, we collect the cross-attention weight Attns $\in \mathbb{R}^{h\times1\times L}$ of each decoder layer to obtain the cross attention weights $\text{Attns}^l = [\text{Attn}^l_1, \text{Attn}^l_2, \text{Attn}^l_3]$ of this step, where h is the number of multi-head and L refers to the size of the flattened image feature. If the token predicted at step l represents a

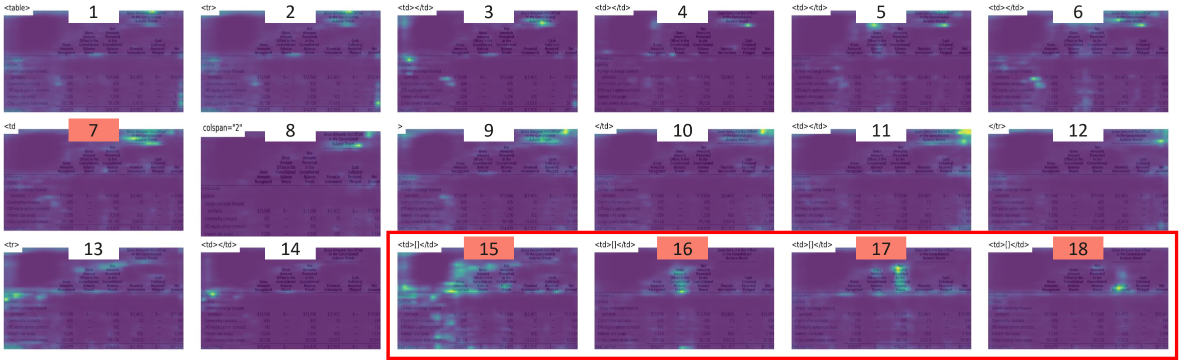


Figure 3. Visualization of cross attention maps of VAST trained **with** visual-alignment loss.

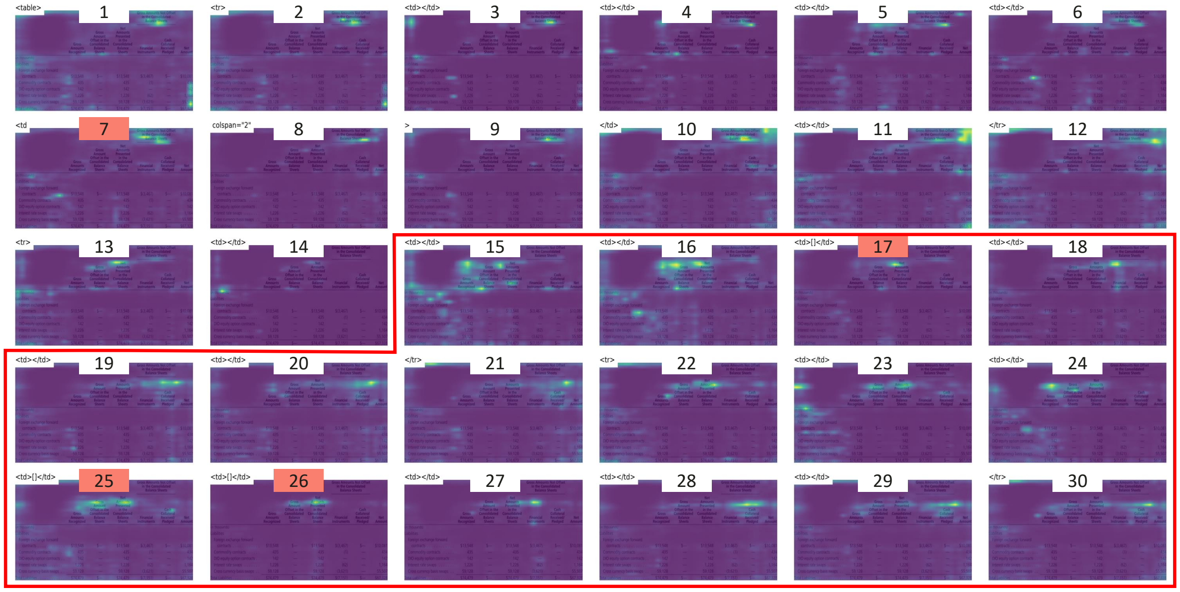


Figure 4. Visualization of cross attention maps of VAST trained **without** visual-alignment loss.

non-empty cell, that is, the token is '<td>[]</td>' or '<td', we average the weights over all layers and all heads to get the averaged weight $\mathrm{Attn}^{l}_{mean} \in \mathbb{R}^{L}$. The averaged weight is reshaped to the size of $\sqrt{L} \times \sqrt{L}$, and the values are normalized to 0-1 and then scaled to 0-255. Finally, the weight map is resized to the size of the image and then overlaid on the original image with 0.8 transparency.

4.2. Comparison of cross-attention visualizations of VAST w/ VA and VAST w/o VA

Fig. 3 and Fig. 4 show cross attention maps of each step as the model predicting the first four non-empty cells. A numeric label with a colored background indicates that the token decoded at this step represents a non-empty cell. Obviously, in the first 14 steps, the logical structure results predicted by the two models are consistent, and the attention maps are also almost similar. The difference occurs in step 15, the model trained with VA loss has more attention near the cell text, so it can correctly predict the cell. However, the model trained without VA loss erroneously focuses on the blank space above the text and incorrectly predicts that cell as a blank cell. After step 15, the difference increases. VAST w/ VA can correctly predict the next three cells at steps 16, 17, and 18, however, due to the error of step 15 and lack of local visual information, VAST w/o VA incorrectly predicts a lot of over-segmented cells.

5. Samples with mutilated columns in SciTSR

Some samples in SciTSR with incomplete columns are shown in Fig. 5. In the table image, it can be seen that the predicted bounding box matches the ground truth accurately except the mutilated columns. In addition, the structure predicted by VAST is also consistent with the image. However, as you can see from the structure of the ground truth, there are some columns that do not exist in the image, which are highlighted by the mask. This results in a lower recall score for our predictions.

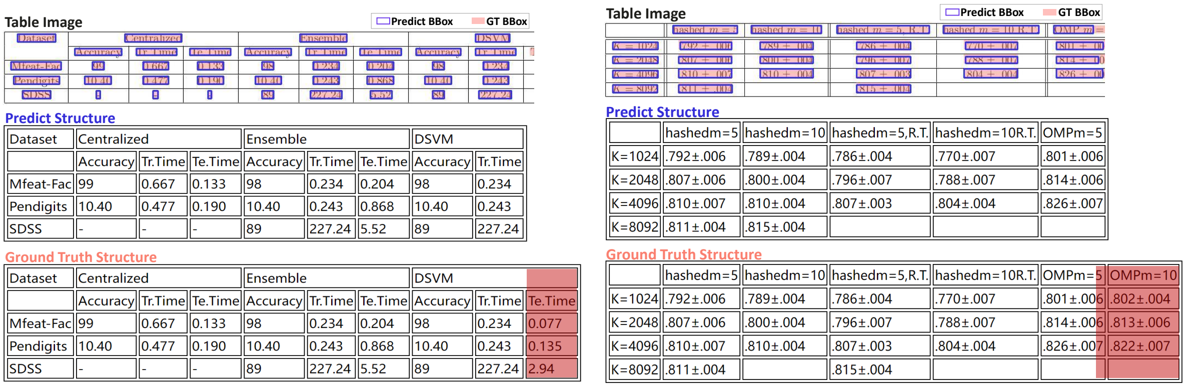


Table Image
Predict BBox
GT BBox
Predict Structure
Ground Truth Structure

Predict Structure

Dataset	Centralized			Ensemble			DSVM	
	Accuracy	Tr.Time	Te.Time	Accuracy	Tr.Time	Te.Time	Accuracy	Tr.Time
Mfeat-Fac	99	0.667	0.133	98	0.234	0.204	98	0.234
Pendigits	10.40	0.477	0.190	10.40	0.243	0.868	10.40	0.243
SDSS	-	-	-	89	227.24	5.52	89	227.24

Ground Truth Structure

Dataset	Centralized			Ensemble			DSVM		
	Accuracy	Tr.Time	Te.Time	Accuracy	Tr.Time	Te.Time	Accuracy	Tr.Time	Te.Time
Mfeat-Fac	99	0.667	0.133	98	0.234	0.204	98	0.234	0.077
Pendigits	10.40	0.477	0.190	10.40	0.243	0.868	10.40	0.243	0.135
SDSS	-	-	-	89	227.24	5.52	89	227.24	2.94

(a) image id:1512.00664v1.7

Predict Structure

	hashedm=5	hashedm=10	hashedm=5,R.T.	hashedm=10R.T.	OMPm=5
K=1024	.792±.006	.789±.004	.786±.004	.770±.007	.801±.006
K=2048	.807±.006	.800±.004	.796±.007	.788±.007	.814±.006
K=4096	.810±.007	.810±.004	.807±.003	.804±.004	.826±.007
K=8092	.811±.004	.815±.004			

Ground Truth Structure

	hashedm=5	hashedm=10	hashedm=5,R.T.	hashedm=10R.T.	OMPm=5	OMPm=10
K=1024	.792±.006	.789±.004	.786±.004	.770±.007	.801±.006	.802±.004
K=2048	.807±.006	.800±.004	.796±.007	.788±.007	.814±.006	.813±.006
K=4096	.810±.007	.810±.004	.807±.003	.804±.004	.826±.007	.822±.007
K=8092	.811±.004		.815±.004			

(b) image id:1202.6384v1.2

Figure 5. Example with mutilated columns in SciTSR. Masked regions in the ground truth are not shown in the image.

References

[1] Yuntian Deng, David Rosenberg, and Gideon Mann. Challenges in end-to-end neural scientific table recognition. In *2019 International Conference on Document Analysis and Recognition*, pages 894–901, 2019. 1

[2] Liangcai Gao, Yilun Huang, Hervé Déjean, Jean-Luc Meunier, Qinqin Yan, Yu Fang, Florian Kleber, and Eva Lang. Icdar 2019 competition on table detection and recognition (ctdar). In *2019 International Conference on Document Analysis and Recognition*, pages 1510–1515, 2019. 1

[3] Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. TableBank: Table benchmark for image-based table detection and recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1918–1925. European Language Resources Association, May 2020. 1

[4] Rujiao Long, Wen Wang, Nan Xue, Feiyu Gao, Zhibo Yang, Yongpan Wang, and Gui-Song Xia. Parsing table structures in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 944–952, October 2021. 1

[5] Sachin Raja, Ajoy Mondal, and C V Jawahar. Visual understanding of complex table structures from document images. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2543–2552. IEEE Computer Society, jan 2022. 1

[6] Asif Shahab, Faisal Shafait, Thomas Kieninger, and Andreas Dengel. An open approach towards the benchmarking of table structure recognition systems. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, DAS '10, page 113–120. Association for Computing Machinery, 2010. 1

[7] X. Zheng, D. Burdick, L. Popa, X. Zhong, and N. Wang. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 697–706. IEEE Computer Society, jan 2021. 1

[8] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: Data, model, and evaluation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 564–580, 2020. 1